Graph-Based Uncertainty Modeling and Multimodal Fusion for Salient Object Detection

Yuqi Xiong[1], Wuzhen Shi[1], Yang Wen[1] (✉), and Ruhan Liu[2]

[1] Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
2022090048@email.szu.edu.cn, wzhshi@szu.edu.cn, wen_yang@szu.edu.cn
[2] Furong Laboratory, Central South University, Changsha, China
223101@csu.edu.cn

Abstract. In view of the problems that existing salient object detection (SOD) methods are prone to losing details, blurring edges, and insufficient fusion of single-modal information in complex scenes, this paper proposes a dynamic uncertainty propagation and multimodal collaborative reasoning network (DUP-MCRNet). Firstly, a dynamic uncertainty graph convolution module (DUGC) is designed to propagate uncertainty between layers through a sparse graph constructed based on spatial semantic distance, and combined with channel adaptive interaction, it effectively improves the detection accuracy of small structures and edge regions. Secondly, a multimodal collaborative fusion strategy (MCF) is proposed, which uses learnable modality gating weights to weightedly fuse the attention maps of RGB, depth, and edge features. It can dynamically adjust the importance of each modality according to different scenes, effectively suppress redundant or interfering information, and strengthen the semantic complementarity and consistency between cross-modalities, thereby improving the ability to identify salient regions under occlusion, weak texture or background interference. Finally, the detection performance at the pixel level and region level is optimized through multi-scale BCE and IoU loss, cross-scale consistency constraints, and uncertainty-guided supervision mechanisms. Extensive experiments show that DUP-MCRNet outperforms various SOD methods on most common benchmark datasets, especially in terms of edge clarity and robustness to complex backgrounds. Our code is publicly available at https://github.com/YukiBear426/DUP-MCRNet.

Keywords: Salient Object Detection · Dynamic Uncertainty · Graph Convolution · Multimodal Fusion · Collaborative Reasoning

1 Introduction

Salient Object Detection (SOD) aims to mimic the human visual system by automatically locating the most attention-grabbing regions in an image. It has broad applications in object recognition [37], semantic segmentation [22], and visual

fixations [24]. Early SOD approaches relied on hand-crafted low-level cues—color contrast, texture, and edges—which work in simple scenes but falter in complex backgrounds due to the absence of high-level semantics.

With the rise of deep learning, CNN-based SOD methods have flourished. DSS [6] introduced deep supervision and short connections within the HED backbone to fuse multi-scale features; Amulet [35] proposed a pyramid fusion strategy to merge spatial details and semantics across levels; PiCANet [12] added an explicit attention module to adaptively model long- and short-range pixel dependencies, enriching local perception. However, fixed-stride down- and up-sampling in these architectures creates an inherent trade-off between preserving high-frequency details and capturing semantic context.

Recently, inspired by Transformers' success in NLP, self-attention has propelled SOD forward. VST [14] was the first to employ a Transformer backbone, capturing long-range dependencies and bolstering global consistency; SRFormer [38] combined Permuted Self-Attention with deep convolutions to mitigate semantic shifts across feature levels; TransformerSOD [16] designed efficient local–global adaptive blocks to balance accuracy and speed. Yet most still use a uniform receptive field, failing to adjust feature complexity per region, which causes blurring or missed detections on fine structures like edges or small objects.

Moreover, most SOD works remain unimodal, overlooking the complementary power of multimodal data in challenging scenes. S2MA [13] and BBSNet [34] show that depth cues compensate for RGB's weaknesses in low-contrast or occluded settings; SAD [1] explores cross-modal reasoning via semantic boundaries and depth guidance, improving robustness. Still, multimodal methods typically process each modality in isolation and lack an effective consistency-driven fusion scheme.

To address these gaps, we propose DUP-MCRNet: a Dynamic Uncertainty Propagation and Multimodal Collaborative Reasoning framework. First, we introduce Dynamic Uncertainty Graph Convolution, which builds a sparse graph based on spatial locality and semantic similarity to propagate uncertainty dynamically across positions, enhancing edge and small-object saliency. Second, we propose a Multimodal Collaborative Fusion strategy that employs learnable modality weights to adaptively fuse attention maps from RGB, depth, and edge features.Our contributions are as follows:

1. **Dynamic Uncertainty Graph Convolution:** Build a sparse graph on spatial–semantic distances to propagate uncertainty across levels, plus channel-wise dynamic fusion to preserve details and semantics.
2. **Multimodal Collaborative Fusion:** Use learnable weights to adaptively combine RGB, depth, and edge features, enhancing cross-modal complementarity and robustness in complex scenes.
3. **Empirical Validation:** Show that DUP-MCRNet consistently outperforms state-of-the-art SOD methods in accuracy and robustness on standard benchmarks.

2 Background

2.1 Multi-scale Feature Fusion

Early salient object detection methods are based on traditional image processing methods such as GC [4] to extract single-scale features or use convolutional neural networks such as DSS [6] to directly predict saliency through shallow or mid-level features. Although these methods can extract local textures, they lack understanding of global structures, resulting in blurred boundaries and poor internal consistency of detection results. In order to overcome the limitation of single scale, Amulet [35] and RA [3] proposed a saliency detection framework based on multi-scale feature aggregation, which fuses features of different resolutions to take into account both fine-grained details and high-level semantic information. However, most of these methods use static splicing or simple layer-by-layer addition, and fail to dynamically adjust the fusion weights according to the uncertainty of spatial position, resulting in insufficient perception of complex boundaries and small objects. PA2Net [32] combines pyramid attention and recursive aggregation strategies to capture salient object information at multiple scales, partially alleviating the problem of spatial information loss during the fusion process. AFNet [5] proposes a feedback mechanism, introduces salient region attention feedback, and guides low-level features to learn more accurate boundary information. However, these methods still use global unified weighting and cannot achieve local adaptive fusion of fuzzy boundary areas, detail-rich areas, and smooth areas. Recently, EGNet [36] uses a boundary guidance module to enhance the boundary perception of salient objects in the feature encoding stage, while BASNet [20] models saliency detection as a residual learning task to gradually refine the details of salient areas. Although these methods have made progress in boundary preservation, they still suffer from boundary detail loss in complex scenarios due to the lack of modeling of feature uncertainty relationships. To this end, this paper proposes a dynamic uncertainty graph convolution and channel adaptive interaction module, which models spatial uncertainty associations through sparse graphs and adaptively guides feature propagation, effectively alleviating the limitations of traditional static fusion methods in fine-grained structure modeling.

2.2 Multimodal Collaborative Reasoning

In order to further improve detection accuracy, researchers have begun to introduce multimodal information to assist saliency detection in recent years. DF [21] and PCF [2] pioneered the exploration of RGB-Depth fusion, using the geometric structure of depth maps to assist saliency reasoning. However, their fusion strategies are mostly simple splicing or early fusion, which fails to fully explore the complementary characteristics between multiple modalities.

CoNet [7] introduced a collaborative attention mechanism, which encodes RGB and Depth features separately and then performs high-order correlation modeling, effectively improving cross-modal reasoning capabilities. DMRA [19]

proposed a multi-scale recursive attention mechanism, which uses depth information to dynamically adjust the attention area of RGB features. In terms of receptive field modeling, traditional methods such as F^3Net [26] and MINet [18] mainly rely on multi-scale feature interactions at a fixed resolution, which makes it difficult to take into account both fine-grained areas and large-scale context modeling. Swin Transformer [15] and PVT [25] introduced local perception capabilities through local window attention and pyramid structure. Inspired by these, this paper proposes a multimodal collaborative fusion, which dynamically adjusts the importance of RGB, depth, and edge features through a learnable modal gating mechanism, effectively improving the saliency detection effect in complex scenes.

3 Methodology

Figure 1 shows the overview architecture of our model. First, the input image is passed through the CNN or Transformer based backbone to extract features of different scales. Then, the features of different levels are passed in turn to the dynamic uncertainty graph convolution and channel adaptive interaction module we proposed to interact with information of different scales. Then, the multimodal collaborative fusion part will use the RGB, Depth and Edge information in the feature map, and perform weighted fusion after self-attention to enhance the model's utilization of multimodal information in the image. Finally, the final output is obtained through an uncertainty enhancement module.

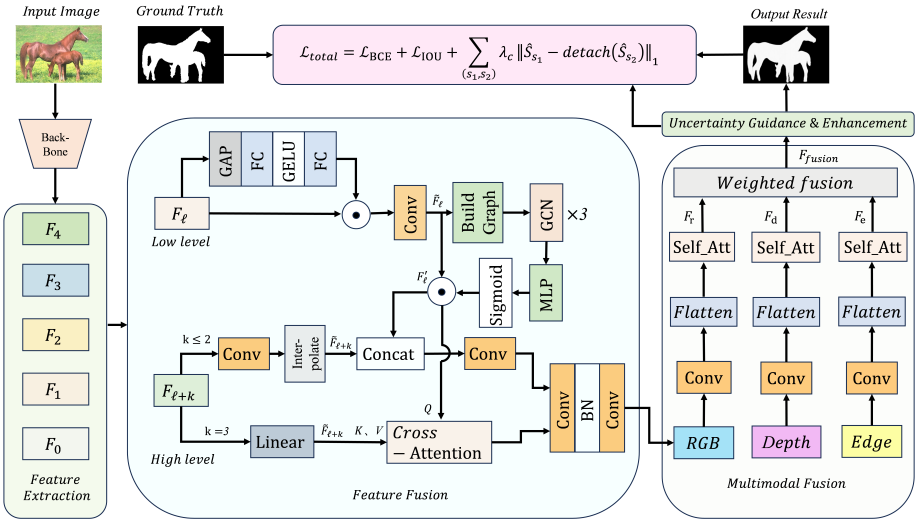

Fig. 1. The overview architecture of our proposed model DUP-MCRNet. Black arrows indicate the data flow. Circle with a dot inside denotes element-wise multiplication.

3.1 Dynamic Uncertainty Graph Interaction

In this module, channel attention is first introduced to the input low-level feature tensor $F_\ell \in \mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, C is the number of channels, representing the feature dimension of each pixel position, H is the height of the feature map, and W is the width of the feature map. It is then mapped to a unified embedding dimension d through a 1×1 convolution to form $\tilde{F}_\ell \in \mathbb{R}^{B\times d\times H\times W}$. Subsequently, $\tilde{F}$ is flattened into a feature sequence of length $N = H \times W$ according to the spatial position and the order is exchanged with the dimensions to obtain $X \in \mathbb{R}^{B\times N\times d}$. Next, a sparse graph is constructed on these spatial nodes for each batch. Specifically, the comprehensive distance $D_{\text{comb}}(i, j)$ is first calculated by a weighted combination of the spatial coordinate distance and the feature similarity, as shown in Formula (1):

$$D_{\text{comb}}(i,j) = \alpha D_{\text{spatial}}(i,j) + (1-\alpha) D_{\text{feature}}(i,j) \tag{1}$$

where $\alpha \in (0, 1)$ is the weight, i,j represents two different spatial positions in the feature map, $D_{\text{spatial}}(i, j)$ is the spatial coordinate distance, and $D_{\text{feature}}(i, j)$ is the feature similarity. The calculation method is shown in Formula (2):

$$D_{\text{spatial}}(i,j) = \|P_i - P_j\|_2 \quad D_{\text{feature}}(i,j) = 1 - \frac{X_i \cdot X_j}{\|X_i\|\|X_j\|} \tag{2}$$

According to $D_{\text{comb}}(i, j)$, the smallest K neighbors are selected for each row to generate the adjacency matrix, as shown in Formula (3):

$$A_{ij} = \begin{cases} 1, & D_{\text{comb}}(i,j) \in \text{TopK}(D_{\text{comb}}(i,:)) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

To ensure self-loop information, we set $\tilde{A} = A + I$ and normalize it with the degree matrix $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. After that, the iterative propagation of uncertainty along the graph structure is explicitly completed through three layers of graph convolution, as shown in Formula (4):

$$X^{(t+1)} = X^{(t)} + \text{ReLU}\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}X^{(t)}W_g^{(t)}\right), t = 0, 1, 2 \tag{4}$$

where $X^{(t)}$ is the feature representation of each node in the t-th graph convolution iteration, and the result is finally reshaped back to the spatial shape $X^{(3)} \in \mathbb{R}^{B\times d\times H\times W}$. In order to highlight the impact of the uncertain region on the subsequent saliency representation, we generate a pixel-level uncertainty weight map through a lightweight MLP and multiply it with $\tilde{F}_\ell$ to obtain $F'_\ell \in \mathbb{R}^{B\times d\times H\times W}$, as shown in Formula (5):

$$F'_\ell = \tilde{F}_\ell \odot \left(1 + \sigma\left(\text{MLP}(X^{(3)})\right)\right) \tag{5}$$

After completing graph convolution and uncertainty enhancement, the module performs channel adaptive compression and spatial alignment on the high-level features $F_{\ell+k}$. For adjacent levels ($k \leq 2$), first reduce its channel to

dimension-d by 1×1 convolution, and then bilinearly interpolate to (H,W), and obtain $\tilde{F}_{\ell+k} \in \mathbb{R}^{B\times d\times H\times W}$. $\tilde{F}_{\ell+k}$ and F'_ℓ are concatenated in the channel dimension, and then two 3×3 convolutions and Relu activation are performed to obtain the fusion feature F_{fuse}; for cross-layer layers ($k = 3$), $F_{\ell+k}$ is channel-projected to obtain $\tilde{F}_{\ell+k} \in \mathbb{R}^{B\times d\times H\times W}$, which is mapped to the Key, $Value$ and $Query$ spaces with the low-level feature $\tilde{F}_\ell$ respectively. Finally, the cross-layer semantic fusion is completed using the cross-attention mechanism, as shown in Formula (6):

$$Attention = \text{Softmax}\left(\frac{(F'_\ell W_q)(\tilde{F}_{\ell+k}W_k)^\mathsf{T}}{\sqrt{d}}\right)\left(\tilde{F}_{\ell+k}W_v\right) \tag{6}$$

W_q, W_k, W_v are learnable parameter matrices. This method uses a cross attention mechanism [29] to fuse features from different levels, using high-level semantics as Key and $Value$ and low-level spatial features as $Query$. This asymmetric attention mechanism exploits the complementarity of scales: global semantics guides the selective extraction of local detail information. Compared with simple concatenation or summation, the cross attention mechanism dynamically weights spatial positions, thereby more accurately fusing semantics and details, enhancing feature discrimination and context awareness.

The obtained attention output is restored to its original shape through linear mapping and then added back to F'_ℓ to complete the interaction, obtaining F_{fuse}. Finally, all cross-layer interaction results are aggregated, and the obtained $F^{(k)}_{\text{fuse}}$ is passed through a set of bottleneck convolutions to obtain the final output $F_{\text{out}} \in \mathbb{R}^{B\times C\times H\times W}$, as shown in Formula (7):

$$F_{\text{out}} = \text{Conv}_{1\times1}\left(\text{RELU}\left(\text{Conv}_{3\times3}\left(F^{(k)}_{\text{fuse}}\right)\right)\right) \tag{7}$$

Through the above design, the model not only explicitly models the uncertainty transmission between different levels in the graph structure, but also tailors the channel fusion strategy for each layer feature, thereby achieving significant improvements in maintaining high-frequency details and cross-layer semantic consistency.

3.2 Multimodal Collaborative Fusion

First, for the cross-modal information fusion part, we perform channel mapping and attention encoding on the three inputs of RGB, Depth and Edge respectively. Assuming that the three modal inputs are ${F_{\text{out}}}^{(m)} \in \mathbb{R}^{B\times C\times H\times W}$, where $m = 1$ represents RGB, $m = 2$ represents Depth, and $m = 3$ represents Edge, multimodal collaborative fusion model maps the m-th modality to a unified embedding dimension through 1×1 convolution, as shown in Formula (8):

$$F^{(m)} = \text{Conv}_{1\times1}\left(F^{(m)}_{\text{out}}\right) \in \mathbb{R}^{B\times d\times H\times W} \tag{8}$$

Then it is flattened according to the spatial dimension to $F^{(m)}_{\text{flat}} \in \mathbb{R}^{B\times d\times N}$, where $N = H \times W$, and then the order of the channel and spatial dimensions

is interchanged to obtain the sequence representation $F_{\text{seq}}^{(m)} \in \mathbb{R}^{B\times N\times d}$. Next, the self-attention mechanism is performed on each modal sequence to obtain the attention output $\hat{F}_{\text{seq}}^{(m)} \in \mathbb{R}^{B\times N\times d}$. Finally, this sequence is reshaped back to the original spatial size $\hat{F}^{(m)} \in \mathbb{R}^{B\times d\times H\times W}$ to complete the global spatial attention encoding of the m-th modality. At the same time, in order to allow the network to automatically learn the importance of each modality, we introduce a learnable modality weight parameter $\theta \in \mathbb{R}^3$ to obtain the weight w_m, and weightedly fuse the attention graphs of each modality, as shown in Formula (9):

$$w_m = \frac{\exp(\theta_m)}{\sum_{k=1}^{3}\exp(\theta_k)} \quad F_{\text{fus}} = \sum_{m=1}^{3} w_m \widehat{F}^{(m)} \tag{9}$$

This fusion method not only retains the complementarity of multi-source information, but also can adaptively adjust the modality weights during training, thereby obtaining more accurate salient area positioning in scenarios such as blurred boundaries and drastic depth changes.

3.3 Loss Function

To effectively guide the model in learning accurate localization and structural detail restoration of salient objects, we design a comprehensive loss function system that integrates hierarchical supervision and fine-grained regularization. The objective is to jointly optimize local pixel-wise accuracy, regional structural consistency, and boundary refinement, thereby enhancing overall prediction quality. The proposed loss function system consists of the following components:

First, to enhance the model's ability to capture multi-level semantics and structural cues, we generate saliency outputs $\{S_{1/16}, S_{1/8}, S_{1/4}\}$ from different decoder stages, corresponding to progressively higher resolutions. These outputs supervise the model in localizing salient regions at multiple scales. In parallel, we design an uncertainty-aware mask prediction branch that produces mask outputs $\{M_{1/4}, M_{1/2}, M_{1/1}\}$, which encode spatial uncertainty and assist in boundary refinement.

For these multi-scale saliency and mask outputs, we employ scale-wise supervision using binary cross-entropy (BCE) and intersection-over-union (IoU) losses. The BCE loss is defined as shown in Formula (10):

$$\mathcal{L}_{\text{BCE}}(\hat{S}, Y) = -\frac{1}{N}\sum_{i=1}^{N}\Big(Y_i \log(\hat{S}_i) + (1-Y_i)\log(1-\hat{S}_i)\Big) \tag{10}$$

where N is the total number of pixels, $\hat{S}$ is the predicted output, and Y is the ground-truth label. To further improve the region-level coherence of the predicted mask, the IoU loss is introduced, as defined in Formula (11):

$$\mathcal{L}_{\text{IoU}}(\hat{S}, Y) = 1 - \frac{\sum_{i=1}^{N}\hat{S}_i Y_i}{\sum_{i=1}^{N}(\hat{S}_i + Y_i - \hat{S}_i Y_i)} \tag{11}$$

Based on these two basic components, the overall saliency loss $\mathcal{L}_{\text{sal}}$ is defined in Formula (12):

$$\begin{aligned}\mathcal{L}_{\text{sal}} = &\sum_{s\in\{1/16,1/8,1/4\}} (\mathcal{L}_{\text{BCE}}(S_s, Y) + \mathcal{L}_{\text{IoU}}(S_s, Y)) \\ &+ \sum_{m\in\{1/4,1/2,1/1\}} (\mathcal{L}_{\text{BCE}}(M_m, Y) + \mathcal{L}_{\text{IoU}}(M_m, Y))\end{aligned} \tag{12}$$

Here, S_s denotes saliency maps at different scales, and M_m denotes the corresponding uncertainty-aware mask outputs. To mitigate potential prediction bias between outputs of different scales, we introduce a cross-scale consistency constraint. For adjacent scale pairs (s_1, s_2), the consistency loss is defined in Formula (13):

$$\mathcal{L}_{\text{consistency}} = \sum_{(s_1,s_2)} \lambda_c \left\| \hat{S}_{s_1} - \text{detach}(\hat{S}_{s_2}) \right\|_1 \tag{13}$$

where λ_c is a weighting coefficient, and $\text{detach}(\cdot)$ prevents gradient propagation to higher-level predictions, ensuring that lower-level features are learned independently. Combining the above losses, our total loss function is expressed as Formula (14):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sal}} + \mathcal{L}_{\text{consistency}} \tag{14}$$

4 Experiment

4.1 Datasets Description

We evaluate our model on six widely used benchmark datasets. SOD [17] contains 300 images from BSD, which contains multiple low-contrast salient objects, often overlapping with image boundaries, and is highly challenging. ECSSD [30] contains 1000 complex scene images, most of which contain a single salient object, with diverse foreground and background patterns, from BSD, VOC2012, and the Internet. PASCAL-S [10] is based on 850 images from PASCAL VOC 2010, and adds eye gaze points and salient segmentation annotations on top of the original annotations. DUT-OMRON [31] contains 5168 images, covering complex backgrounds and multiple salient objects, and provides accurate pixel-level annotations. HKU-IS [9] contains 4447 images with salient object annotations, and the images meet at least one of the following conditions: there are multiple unconnected objects, objects touch edges, or the color contrast is less than 0.7. DUTS [23] is the largest saliency detection dataset currently, containing 10,553 training images and 5,019 test images, selected from the ImageNet and SUN datasets respectively. All annotations are manually completed by 50 participants.

4.2 Experimental Setup

We use 10,553 images from the DUTS dataset for training, and resize all images to 384×384. The batch size is set to 8, and the model is trained for 100 epochs. The Adam optimizer [8] is employed with an initial learning rate of 1×10^{-4}, and a learning rate decay strategy is applied to facilitate convergence. All experiments are conducted on an NVIDIA RTX 4090 GPU with 24 GB memory.

To comprehensively evaluate the performance of the proposed method, we adopt four mainstream evaluation metrics: mean absolute error (MAE), mean intersection-over-union (mSIOU), structural similarity measure (S-measure), and weighted F-measure (Weighted-F_β). The specific definitions of these metrics are provided below:

Mean Absolute Error (MAE): MAE measures the average absolute pixel-wise difference between the predicted saliency map P and the ground truth label G, and is defined as shown in Formula (15):

$$\mathrm{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |P(x,y) - G(x,y)| \tag{15}$$

where H and W denote the height and width of the image, respectively. $P(x,y)$ and $G(x,y)$ represent the predicted and ground truth values at pixel (x,y). A smaller MAE indicates better prediction accuracy.

Mean Structural IoU (mSIOU): mSIOU evaluates the structural overlap between predicted regions and ground truth, reflecting the consistency at a structural level. It is computed as shown in Formula (16):

$$\mathrm{mSIOU} = \frac{1}{N} \sum_{i=1}^{N} \frac{P_i \cap G_i}{P_i \cup G_i} \tag{16}$$

In this equation, P_i and G_i denote the predicted and ground truth binary regions for the i-th instance, and N is the total number of pixels or segments. Higher mSIOU values indicate better structural alignment.

Structural Similarity Measure (S-measure): S-measure combines object-level and region-level structural similarities into a unified metric. It is defined as shown in Formula (17):

$$S_m = \alpha S_o + (1 - \alpha) S_r \tag{17}$$

where S_o is the object-level structural similarity, S_r is the region-level similarity, and α is a balancing coefficient. A higher S_m value implies that both global and local structures are well preserved.

Weighted F-measure (Weighted-F_β): This metric extends the traditional F-measure by incorporating pixel-wise weights in the computation of precision and recall, which allows better handling of imbalanced and spatially variant errors. It is defined in Formula (18):

$$F_\beta^w = \frac{(1+\beta^2)\cdot \text{Precision}_w \cdot \text{Recall}_w}{\beta^2 \cdot \text{Precision}_w + \text{Recall}_w} \tag{18}$$

Here, β^2 adjusts the relative importance of precision and recall. Weighted-F_β better captures the perceptual quality of predictions by considering both spatial and neighborhood error distributions, especially for salient objects with blurred or incomplete boundaries.

4.3 Experimental Results

Table 1 shows the quantitative comparison results on five benchmark datasets. We compare our method against five state-of-the-art models: CPD [28], PoolNet [11], LDF [27], MINet [18], and UGRAN [33]. The four indicators of *MAE*, *mSIOU*, S_m, and F_β^w are used to evaluate the performance of six methods. The results show that our method outperforms previous advanced methods in the indicators of most datasets.

Table 1. Comparison results of different methods on five benchmark datasets. Bold indicates the best result and underline indicates the second best result.

Dataset	DUT-O				DUTS				ECSSD				HKU-IS				PACSAL-S			
Metrics	$MAE\downarrow$	$mS\uparrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$mS\uparrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$mS\uparrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$mS\uparrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$mS\uparrow$	$S_m\uparrow$	$F_\beta^w\uparrow$
CPD	0.057	0.741	0.818	0.715	0.043	0.795	0.867	0.800	0.040	0.839	0.910	0.895	0.033	0.831	0.904	0.879	0.072	0.748	0.845	0.796
PoolNet	0.056	0.754	0.836	0.729	0.040	0.807	0.883	0.807	0.039	0.851	0.921	0.896	0.032	0.850	0.917	0.883	0.075	0.766	0.849	0.723
LDF	**0.052**	0.772	0.839	0.752	**0.034**	0.828	0.892	0.845	0.034	0.861	0.924	0.915	0.028	0.857	0.919	0.904	**0.051**	**0.801**	**0.882**	0.847
MINet	0.057	0.753	0.822	0.718	0.039	0.804	0.875	0.813	0.036	0.857	0.919	0.905	0.031	0.853	0.912	0.889	0.064	0.763	0.854	0.808
UGRAN	0.058	0.768	0.830	0.733	**0.034**	**0.856**	**0.924**	**0.911**	0.034	0.856	0.924	0.911	0.028	0.856	0.922	0.905	0.059	0.773	0.867	0.826
Ours	0.066	**0.788**	**0.840**	**0.755**	0.036	0.845	0.902	0.858	**0.030**	**0.881**	**0.935**	**0.922**	**0.027**	**0.876**	**0.929**	**0.913**	0.061	0.786	0.869	**0.930**

In Figure 2, we plot the precision-recall curves of each method to comprehensively evaluate the detection performance of the model at different thresholds. From the results, it can be seen that our method achieves a high balance between precision and recall, which verifies the superior performance and good robustness of the proposed model in the task of salient object detection.

Figure 3 provides a visual comparison of our model with other methods. It can be seen that our method has higher accuracy in restoring saliency maps and effectively reduces the interference of shadows and low-saturation areas. In addition, it shows better integrity and robustness when there are complex backgrounds or detailed structure maps.

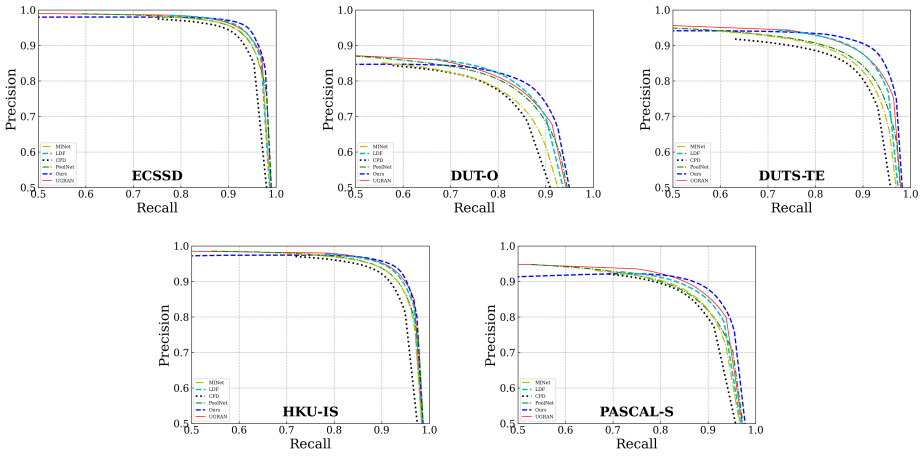

Fig. 2. Precision-recall curves of different methods on the salient object detection task. The blue dashed line represents our model, and the other lines represent the models we compare against.

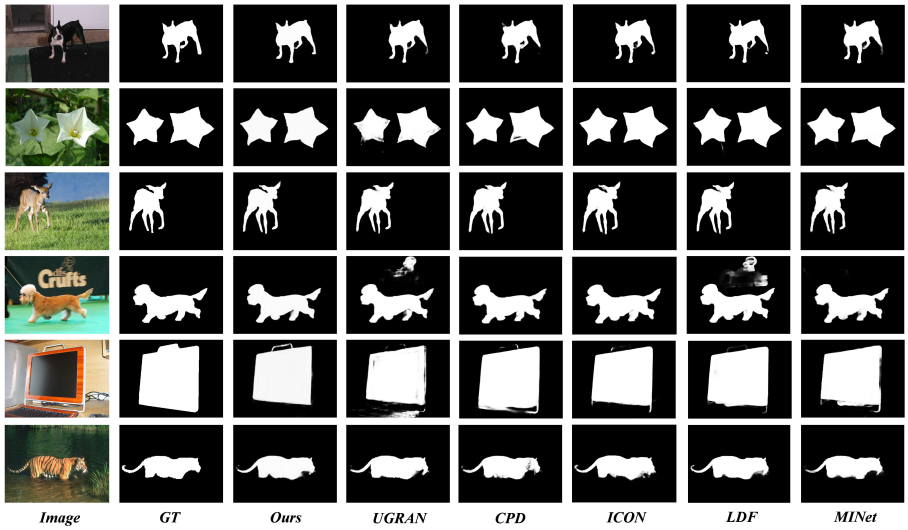

Fig. 3. Visualization of comparative results for various saliency detection models. Our method better preserves salient structures and suppresses noise from shadows and low-saturation regions.

4.4 Ablation Study

In order to further verify the effectiveness of each module, we use the basic model as a comparison, and gradually introduce DUGC and MCF modules to observe the effect of each module on the overall performance improvement. The experimental results are shown in Table 2, which shows that the introduction of each module has a positive effect on the performance improvement, verifying the effectiveness and necessity of the proposed design.

Table 2. Ablation study results across different datasets. Bold indicates the best result.

Datasets			ECSSD				HKU-IS				SOD			
Base	DUGC	MCF	$MAE \downarrow$	$mS \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$mS \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$mS \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$
✓	✗	✗	0.034	0.856	0.924	0.911	0.028	0.856	0.922	0.905	0.094	0.665	0.795	0.758
✗	✓	✗	**0.028**	0.858	0.922	0.905	0.028	0.858	0.922	0.905	0.090	0.677	0.801	0.765
✗	✗	✓	0.031	0.872	0.930	0.915	**0.026**	0.873	0.928	0.912	0.082	0.697	0.818	0.780
✗	✓	✓	0.030	**0.881**	**0.935**	**0.922**	0.027	**0.876**	**0.929**	**0.913**	**0.078**	**0.711**	**0.827**	**0.796**

Specifically, when the DUGC module is introduced alone, most of the indicators of the model on the three datasets are improved, indicating that the dynamic modeling uncertainty graph helps capture the complex relationships in the data. Similarly, the MCF module is introduced alone to bring performance improvements, which proves the effectiveness of multimodal collaborative fusion in enhancing feature representation and robustness.

More importantly, combining the DUGC and MCF modules achieves the best results on almost all indicators and datasets, indicating the complementary effects between the two components. This synergy enables the model to better utilize uncertainty propagation and multimodal information, resulting in more accurate and stable predictions.

5 Limitation

Although the proposed DUP-MCRNet achieves strong performance on multimodal saliency detection benchmarks, there remain limitations worth further exploration. First, the DUGC module introduces graph construction and multi-step uncertainty propagation, which, despite enhancing feature representation, incur high computational and memory overhead—especially with high-resolution inputs or multi-frame sequences—hindering real-time use on resource-limited platforms. Second, cross-domain generalization has yet to be fully validated. Existing experiments focus on standard datasets, but performance under real-world conditions—such as nighttime infrared or adverse weather—remains unclear, calling for further robustness evaluation.

6 Conclusion

This paper introduces DUP-MCRNet, a novel framework for salient object detection that tackles detail loss, edge ambiguity, and suboptimal multimodal fusion in complex visual scenes. The proposed Dynamic Uncertainty Graph Convolution (DUGC) module explicitly models uncertainty propagation via sparse spatial-semantic graphs, enabling adaptive refinement of small-scale structures and ambiguous boundaries. In parallel, the Multimodal Collaborative Fusion (MCF) strategy leverages learnable gating weights to dynamically integrate RGB, depth, and edge features, facilitating coherent global-local reasoning across modalities. Extensive experiments on five public benchmarks demonstrate that DUP-MCRNet consistently outperforms most of the state-of-the-art methods, especially in preserving edge sharpness and maintaining robustness in cluttered backgrounds. Ablation studies further validate the complementary strengths of the DUGC and MCF modules, showing their synergy significantly enhances boundary preservation and adaptability to challenging scenes. In future work, we aim to explore lightweight model deployment and extend the proposed uncertainty-aware fusion mechanisms to video-based saliency detection tasks.

References

1. Cen, J., Wu, Y., Wang, K., Li, X., Yang, J., Pei, Y., Kong, L., Liu, Z., Chen, Q.: Sad: Segment any rgbd (2023), https://arxiv.org/abs/2305.14207
2. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3051–3060 (2018). https://doi.org/10.1109/CVPR.2018.00322
3. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection (2019), https://arxiv.org/abs/1807.09940
4. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2015). https://doi.org/10.1109/TPAMI.2014.2345401
5. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1623–1632 (2019). https://doi.org/10.1109/CVPR.2019.00172
6. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(4), 815–828 (Apr 2019). https://doi.org/10.1109/tpami.2018.2815688, http://dx.doi.org/10.1109/TPAMI.2018.2815688
7. Ji, W., Li, J., Zhang, M., Piao, Y., Lu, H.: Accurate rgb-d salient object detection via collaborative learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 52–69. Springer International Publishing, Cham (2020)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), https://arxiv.org/abs/1412.6980

9. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5455–5463 (2015). https://doi.org/10.1109/CVPR.2015.7299184
10. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation (2014), https://arxiv.org/abs/1406.2807
11. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3912–3921 (2019). https://doi.org/10.1109/CVPR.2019.00404
12. Liu, N., Han, J., Yang, M.H.: Picanet: Pixel-wise contextual attention learning for accurate saliency detection. IEEE Transactions on Image Processing **29**, 6438–6451 (2020). https://doi.org/10.1109/TIP.2020.2988568
13. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13753–13762 (2020). https://doi.org/10.1109/CVPR42600.2020.01377
14. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer (2021), https://arxiv.org/abs/2104.12099
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
16. Mao, Y., Zhang, J., Wan, Z., Tian, X., Li, A., Lv, Y., Dai, Y.: Generative transformer for accurate and reliable salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **35**(2), 1041–1054 (2025). https://doi.org/10.1109/TCSVT.2024.3469286
17. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. pp. 49–56 (2010). https://doi.org/10.1109/CVPRW.2010.5543739
18. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection (2020), https://arxiv.org/abs/2007.09062
19. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7253–7262 (2019). https://doi.org/10.1109/ICCV.2019.00735
20. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7471–7481 (2019). https://doi.org/10.1109/CVPR.2019.00766
21. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing **26**(5), 2274–2285 (May 2017). https://doi.org/10.1109/tip.2017.2682981, http://dx.doi.org/10.1109/TIP.2017.2682981
22. Sun, G., Wang, W., Dai, J., Gool, L.V.: Mining cross-image semantics for weakly supervised semantic segmentation (2020), https://arxiv.org/abs/2007.01947
23. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3796–3805 (2017). https://doi.org/10.1109/CVPR.2017.404

24. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey (2021), https://arxiv.org/abs/1904.09146
25. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions (2021), https://arxiv.org/abs/2102.12122
26. Wei, J., Wang, S., Huang, Q.: F3net: Fusion, feedback and focus for salient object detection (2019), https://arxiv.org/abs/1911.11445
27. Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection (2020), https://arxiv.org/abs/2008.11048
28. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection (2019), https://arxiv.org/abs/1904.08739
29. Xiong, Y., Wen, Y.: Non-stationary time series forecasting based on fourier analysis and cross attention mechanism (2025), https://arxiv.org/abs/2505.06917
30. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1155–1162 (2013). https://doi.org/10.1109/CVPR.2013.153
31. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3166–3173 (2013). https://doi.org/10.1109/CVPR.2013.407
32. Yu, J., Liu, Y., Wu, X., Xu, K., Li, J.: Pa2net: Pyramid attention aggregation network for saliency detection. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 186–200. Springer Nature Singapore, Singapore (2025)
33. Yuan, Y., Gao, P., Dai, Q., Qin, J., Xiang, W.: Uncertainty-guided refinement for fine-grained salient object detection. IEEE Transactions on Image Processing **34**, 2301–2314 (2025). https://doi.org/10.1109/TIP.2025.3557562
34. Zhai, Y., Fan, D.P., Yang, J., Borji, A., Shao, L., Han, J., Wang, L.: Bifurcated backbone strategy for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 8727–8742 (2021). https://doi.org/10.1109/tip.2021.3116793, http://dx.doi.org/10.1109/TIP.2021.3116793
35. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 202–211 (2017). https://doi.org/10.1109/ICCV.2017.31
36. Zhao, J., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8778–8787 (2019). https://doi.org/10.1109/ICCV.2019.00887
37. Zhou, T., Fan, D.P., Cheng, M.M., Shen, J., Shao, L.: Rgb-d salient object detection: A survey. Computational Visual Media **7**(1), 37–69 (Mar 2021). https://doi.org/10.1007/s41095-020-0199-z, http://dx.doi.org/10.1007/s41095-020-0199-z
38. Zhou, Y., Li, Z., Guo, C.L., Bai, S., Cheng, M.M., Hou, Q.: Srformer: Permuted self-attention for single image super-resolution. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12734–12745 (2023). https://doi.org/10.1109/ICCV51070.2023.01174